"AFFINITY IN PAIRED-EVENT PROBABILITY"

by

S.P. Blythe*†, S. Busenberg‡ and C. Castillo-Chavez†

*Department of Statistics and Modelling Science,
Strathclyde University,
Livingstone Tower,
Glasgow G1 1XH,
Scotland.

†Biometrics Unit,
337 Warren Hall,
Cornell University,
Ithaca, NY 14853
U.S.A.

‡Department of Mathematics,
Harvey Mudd College,
Claremont, California 91711
U.S.A.

ABSTRACT

Some recent advances in sexually transmitted disease epidemiology have a wider applicibility in applied probability. A general parametric functional may generate the conditional and joint probabilities of event-pairs where the within-pair event order is irrelevant. The parameters (ϕ) represent affinities or associations between single events, and if the marginal probabilities of the single events are known, then the ϕ specify a hyper-surface on which all the joint probabilities of event-pairs must lie. A proof is given, some examples are presented, and applications in ecology, epidemiology, and distribution theory are offered.

INTRODUCTION

In the epidemiology of sexually transmitted diseases, the problem of modelling arbitrary mixing patterns of distinct population sub-groups (differentiated by *eg* sex, number of partners, age) has in the last few years received a great deal of attention, due to the implications of AIDS as a sexually transmitted disease. Recently, a general solution to the problem has been found (Busenberg and Castillo-Chavez, 1989, 1990), structurally taking the form of parameterized multiplicative perturbations of random mixing, the non-randomness arising from the preference or affinity between particular sub-groups. As sexual mixing is essentially a paired-event process, it seems reasonable to express the problem and the solution in general probabalistic terms, in order to bring the work to the attention of a larger section of the community, and to illustrate that a wider applicability than the original mixing problem exists.

We show that the axiomatic framework underlying the thereoms of Busenberg and Castillo-Chavez (1989, 1990) is equivalent to specifying a process where two elementary events occur at each "event-pair", with the order of elementary events within a pair being immaterial. Hence we show that a parameterized representation of all such processes may be stated explicitly. We use a number of examples to illustrate the form and structure of the paired-event representation, and indicate a few areas where it is applicable. We conclude with a brief discussion of these applications, and offer suggestions for future work.

PAIRED EVENT PROBABILITY

Let $S = \{i\ ;\ 1 = 1,2,\ldots,\ n\}$ be an index set, and let $\Omega = \{E_i\ ;\ i \in S\}$ be the finite sample space of disjoint elementary events E_i. Let $A \in \Omega$ be the event-space of un-ordered 2-tuples, $A = \{E_i \cap E_j\ ;\ i \in S,\ j \in S\}$, containing n^2 elements. For convenience we index the elements of A as $A_k = (X = E_i \cap Y = E_j)$, where (X, Y) is the event pair ***first*** and ***second***, and $k = (i-1) \times n + j$ (*i.e.* vectorizing the $n \times n$ matrix). We will denote by

$\rho_i \equiv \Pr\{E_i\},\ i \in S,\ E_i \in \Omega,$ the marginal probabilities for events;

$p_{ij} \equiv \Pr\{X = E_i \mid Y = E_j\}$, $(i,j) \in S$, $(E_i, E_j) \in \Omega$, the conditional probabilities for event-pairs;

$\pi_{ij} \equiv \Pr\{X = E_i \cap Y = E_j\}$, $(i,j) \in S$, $(E_i, E_j) \in \Omega$, the joint probabilities of event-pairs.

The following trivial restrictions to the problem are necessary:

$$\rho_i \in [0, 1], \quad i \in S \tag{1}$$

$$\sum_{i \in S} \rho_i = 1 \tag{2}$$

and

$$\rho_i = 0 \ \vee \rho_j = 0 \ \Rightarrow \ p_{ij} = p_{ji} = \pi_{ij} = \pi_{ji} = 0 \ , \tag{3}$$

reflecting the fact that the ρ are probabilities, and that non-existent events cannot occur in event-pairs. Let $S' \in S$ be the subset where $\rho_i > 0$. We impose the non-trivial restriction

$$\pi_{ij} = \pi_{ji} \ , \quad (i, j) \in S', \tag{4}$$

which specifies the class of problems to be addressed: those where the order of events within an event-pair is immaterial. Note that Eq (4) is *not* the same as $p_{ji} = p_{ij}$. Then we have the following:-

Theorem 1

For a given marginal distribution of elementary events ρ, $\exists$ an $n \times n$ matrix of constants ϕ, with $\phi_{ij} = \phi_{ji} \geq 0$, such that

$$\pi_{ij} = \rho_i \, \rho_j \left[\frac{R_i \, R_j}{\sum_{m \in S} \rho_m \, R_m} + \phi_{ij} \right] , \ (i,j) \in S' \tag{5}$$

where

$$R_i = \sum_{m \in S} \rho_m \, (1 - \phi_{im}) \ , \ i \in S'. \tag{6}$$

The ϕ are constrained by $R_i \geq 0$, $i \in S'$, with at least one $R_i > 0$.

Proof By the definition of conditional probability, Eq (4) implies that

$$p_{ij}\,\rho_j = p_{ji}\,\rho_i\ , \quad (i,j) \in S, \tag{7}$$

so that Eqns (1)−(4) constitute the axiomatic framework used by Busenberg and Castillo-Chavez (1989, 1990), and so their Representation Theorem for p holds (their Thereom 4.4), for ϕ satisfying the hypotheses of Thereom 1, and R_i given by Eq (6), and the proof is complete. In Appendix A we give a brief outline of the Busenberg and Castillo-Chavez (1990) proof.

Remark The parameters ϕ are a measure of affinity of each kend of event for every other kind. If $\phi_{ij} = x \geq 0$, $(i,j) \in S'$, then (9) reduces to

$$\pi_{ij} = \rho_i\,\rho_j\ , \tag{8}$$

which is the familiar result for independent events. In fact, it may be shown the independence result Eqn (8) arises for $0 \leq \phi_{ij} \leq 1$, $(i,j) \in S$, iff the elements of ϕ are given by

$$\phi_{ij} = 1 - \eta_i\,\eta_j\ ,$$

where $0 \leq \eta_i \leq 1$, all i (see Blythe, *in prep.*). Different choices of ϕ produce all the possible joint distributions of events in pairs, where within-pair order is immaterial. If all event-pairs in A can occur, there are $n(n+1)/2$ independent ϕ_{ij} values; where some pairings are "forbidden" (other than in the sense of Eq (3)), Eq (5) and (6) are defined over the permissible event-pairs (*c.f.* Appendix A; Busenberg and Castillo-Chavez 1989, 1990; Blythe 1991).

EXAMPLES AND APPLICATIONS

Example 1 : Color-Bias in Card-Games

In a trivial version of the card game patience (solitaire), a game consists of shuffling the deck, and dealing out the top two cards. If these are of the same color the player wins, otherwise (s)he loses. When the deck is fair, then 2 reds occur with probability 0.25, 2 blacks with probability 0.25, and two

mis-matched cards with probability 0.5. For an *unfair* deck, where cards have some tendency to remain together through the shuffle, on the basis of color alone, the outcomes will differ from the random result. We may use the ϕ-method to examine the problem of "affinities" between cards of different colors.

Here we have two events: E_1 = "red card", and E_2 = "black card", with $\rho_1 = \rho_2 = \frac{1}{2}$. If we write

$$\phi = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \tag{9}$$

then we have at once from Eq (5) the matrix of joint probabilities for all possible event-pairs,

$$\pi = \begin{bmatrix} \frac{1}{4}\left[\frac{[1-\frac{1}{2}(a+b)]^2}{1-\frac{1}{4}(a+2b+c)} + a\right] & \frac{1}{4}\left[\frac{[1-\frac{1}{2}(a+b)][1-\frac{1}{2}(b+c)]}{1-\frac{1}{4}(a+2b+c)} + b\right] \\ \frac{1}{4}\left[\frac{[1-\frac{1}{2}(a+b)][1-\frac{1}{2}(b+c)]}{1-\frac{1}{4}(a+2b+c)} + b\right] & \frac{1}{4}\left[\frac{[1-\frac{1}{2}(b+c)]^2}{1-\frac{1}{4}(a+2b+c)} + c\right] \end{bmatrix}. \tag{10}$$

For example, we have

$$\phi = \begin{bmatrix} 2 & 0 \\ 0 & c \end{bmatrix} \Rightarrow \pi = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \tag{11}$$

$$\phi = \begin{bmatrix} a & a \\ a & a \end{bmatrix} \Rightarrow \pi = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \tag{12}$$

and

$$\phi = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix} \Rightarrow \pi = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}. \tag{13}$$

This example gives a good illustration of the value of the ϕ-method. Consider the ϕ-values as a measure of color-for-color affinity among cards, brought about by the shuffling process; $\phi_{ij} = 0$ implies no affinity, while $\phi_{ij} = 2$ (in this example) implies maximum affinity. If all affinities are equal, the result is a random pattern (Eq (12)). If just one color has maximum mutual affinity, and there are zero inter-color affinities, then cards of that color are always found together, so that color-separation takes place, and the result is a guaranteed win, regardless of the self-affinitiy of the other color (Eq (11)). To get a guaranteed loss (no same-color pairs, Eq (13)), the inter-color affinity must be maximum, and both self-affinities zero.

"Affinity" here really means that the shuffling process favors or disfavors like-wth-like color matches, making cards adjacent with a probability greater than random. The ϕ thus measure bias in the shuffle, and provide a way of quantifying the degree of certain kinds of non-randomness. It is interesting to note that bias (cheating) which causes a *particular* color always to win involves changing the marginals, setting say $\rho_1 = 0$ for black to win with probability one.

It may be possible to use the ϕ-method in the analysis of a variety of paired-event experiments, perhaps in the sense discussed by Mielke and Siddiqui (1965), where independence is normally "expected", but where outside influences sometimes correlated occurences (in their case, the presence of atmospheric pollutants caused non-random temporal clustering of asthmatic attacks among three subjects).

Example 2: Ecological Association in Plant Communities

A problem of long-standing interest in plant ecology is the degree to which species tend to be found together in particular habitats (see Ludwig and Reynolds 1988, the source for much of the following background information). Legendre and Legendre (1983) refer to this kind of study as *R-mode analysis.* One particular set of questions concerns inter-specific association (also referred to in the literature as affinity), arising because *(a)* both species select or avoid the same kinds of habitat, *(b)* both species have approximately the same requirements, or *(c)* there is a genuine affinity or disaffinity between them (Hubalek, 1982; Ludwig and Reynolds 1988). Ludwig and Reynolds (1988, Ch 11) discuss this at length, and present many practical schemes for detecting and classifying associations. An

assortment of indices, measures, and statistical tests have been developed for studying inter-species associations, some of which work quite effectively when tested against known, artificial, data (*op. cit.*). Both presence-absence indices (association) and those based on abundance (covariation) have been developed. All these measures are essentially *ad hoc* in nature, a general theory of asociations being conspicuous by its absence. We speculate that the paired-event affinity structure of Eq (5) may provide a useful framework for investigating association. The key result which remains to be obtained is the elaboration of precise experimental protocols for sampling. We speculate that an experimental design capturing the essential unordered event-pair characteristic of the method would involve sampling two linked points in an appropriate sampling unit: perhaps by looking at the two nearest plants to each of a series of randomly chosen points in a unit, and replicating over many units to get the multiplicity of sampled joint distributions required uniquely to specify the elements of ϕ.

If data on π can be obtained for several (say M) replicates, then we can estimate the ϕ-values directly, using for example the method of least-squares. There are n(n+1)/2 independent ϕ-values, but only n(n-1)/2 independent values in a p or π matrix, so that we require at least n+1 samples to fit the ϕ.

As a preliminary example of how the ϕ-method might be applied, we choose a published data set where we can approximately (but *only* approximately) extract the information we need to use the method, namely that of Krebs (1978, p. 375 *et seq.*). An ecology class looked at the association between two species of grasses on an area of sand-dunes on the southern shores of Lake Michigan. They performed both quadrat presence-absence and nearest-neighbor analyses. The species were *Andropogon scoparius* (Species 1 here), and *Ammophilia breviligulata* (Species 2 here). The data are not ideal for our purposes, but we can extract two rough estimates of π, one from the presence-absence data, and one from the nearest-neighbor data. With E_1 = "Species 1 found", and E_2 = "Species 2 found", and using the notation $\pi^{(k)}$ for the k^{th} data matrix (with M = 2), we have

$$\pi^{(1)} = \begin{bmatrix} 0.3615 & 0.0308 \\ 0.0308 & 0.5769 \end{bmatrix}, \Rightarrow \rho^{(1)} = \begin{bmatrix} 0.3923 \\ 0.6077 \end{bmatrix}, \tag{17}$$

and

$$\pi^{(2)} = \begin{bmatrix} 0.3214 & 0.0863 \\ 0.0863 & 0.5060 \end{bmatrix}, \Rightarrow \rho^{(2)} = \begin{bmatrix} 0.4077 \\ 0.5923 \end{bmatrix}. \tag{18}$$

Thus we have two (admittedly questionable) estimates of π for the dune-grasses. Now note that in this n = 2 case, only one element in π, for each k, is independent, but that there are 3 independent ϕ-values. Although this means we cannot uniquely specify the ϕ-values form the available data, we can however test the fit of two contradictory hypotheses, namely that (A) intra-species affinity dominates, and (B) inter-species affinity dominates. We can express these hypotheses using test matrices for ϕ,

$$\mathcal{H}_A: \qquad \phi = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \tag{14}$$

and

$$\mathcal{H}_B: \qquad \phi = \begin{bmatrix} 0 & b \\ b & 0 \end{bmatrix}. \tag{15}$$

Of course, many other one-parameter families may be tested (*e.g.* all distinct combinations of 0, 1, and a parameter): the point here is to show that it can be done, not to perform an exhaustive analysis. We estimate the parameters in Eq (14) or (15) using the obvious scheme, minimising

$$S(\phi) = \sum_{k=1}^{M} \sum_{i=1}^{n-1} \sum_{j=i}^{n-1} \left(\pi_{ij}^{(k)} - \hat{\pi}_{ij} \right)^2 , \tag{16}$$

where $\hat{\pi}_{ij}$ is the predicted value from Eq (5) using ϕ from Eq (14) or (15). The outer sum is over replicates, the next over rows in π, and the inner sum over columns; there are n(n-1)/2 terms, equalling the number of independent π_{ij}. Using the inter-species affinity model, Eq (15), we get

$$\phi = \begin{bmatrix} 1.425 & 0 \\ 0 & 1.425 \end{bmatrix}, \tag{17}$$

while from Eq (15) we can only obtain

$$\phi = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{18}$$

i.e. random interactions. The corresponding R^2 values for each fit are approximately 0.41 and 0.12, respectively: neither is wonderful, but $\mathcal{H}_A$ is clearly a lot better than $\mathcal{H}_B$. Given that $\mathcal{H}_B$ can only manage random effects with the given structure, we can feel reasonably sure that the species dis-

associate. Without further data in this case, we can say little more, but the conclusion agrees with those of the standard tests discussed by Krebs (1978).

Example 9: Single-sex mixing models

(a) In its original form, the result of Busenberg and Castillo-Chavez (1989, 1990) deals with sexual contacts (*i.e.* partners) per unit time in a population comprised of n groups, in the i^{th} of which there are at time t $T_i(t)$ individuals with average number of partners per unit time, or "risk", $C_i(t)$. A valid description of the mixing process is produced by an $n \times n$ matrix of probabilities p(t) where $p_{ij}(t)$ is the probability that an individual in group i has a partner in group j, at time t. Constraints (1) and (2) simply make p a stochastic matrix, (4) enforces conservation of the number of new pairings per unit time between individuals in groups, and (3) says that individuals in momentarily empty or inactive groups cannot have partners.

The problem of developing deterministic p(t), functions of C(t) and T(t) only, which satisfy the axiomatic constraints, has become one of extreme importance in modeling the epidemiology of sexually-transmitted diseases (STDs), and in particular the human immunodeficiency virus (HIV, the causative agent of AIDS). Unitl recently, the only known solution was random or "proportionate" mixing (*e.g.* Barbour 1978, Nold 1980, Hethcote and Yorke 1984, Anderson and May 1984, Dietz and Schenzle 1985, Anderson *et al* 1987, Blythe and Anderson 1989, Castillo-Chavez *et al.* 1988, 1989*a*), where $p_{ij} = \rho_j$, all i, with in this context the marginals being given by

$$\rho_j(t) = \frac{C_j(t)T_j(t)}{\sum_{k=1}^{n} C_k(t)T_k(t)} \quad , \text{all } j \ . \tag{19}$$

In addition, a version of the assortative mating structure familiar to population geneticists has been used in the STD literature (where it is called "preferred mixing"; *e.g.* Nold 1980, Hethcote and Yorke 1984, Jacquez *et al.* 1988, 1989), and models with rule-based adaptive sexual behavior have been suggested by Anderson *et al.* (1989, 1990) and Gupta *et al.* (1989). A number of particular-case "mixing functions" have also recently been proposed (e.g. Castillo-Chavez and Blythe 1989; Hyman and Stanley 1989; Koopman *et al.* 1988, 1989), but the field has been significantly opened up by Busenberg and Castillo-Chavez (1989, 1990) who generalized the specific case of Blythe and Castillo-Chavez (1989) – called "like-with-like mixing" – and obtained the representation theorem (see Appendix A) stating that all mixing functions may be expressed in a special form. Blythe and Castillo-

Chavez 1990, and Blythe *et al.* (1991*b*) show how a selection of particular cases fit into the general form, and the latter paper also discusses the implementation of mixing and pair formation/dissolution models. Specific two-sex, age-structured, multi-partner results have also been obtained (Castillo-Chavez and Busenberg, 1990) as well as solutions for arbitrarily connected groups of any type (Blythe, 1991).

The representation theorem states that *any* p, a solution to the problem specified by constraints Eqs (1) – (4), may be written in the form given by Eq (5). The parameters ϕ (which may in this context be functions of $\rho(t)$), provide a measure of *mutual preference*, or *affinity* for sexual partners between pairs of groups (see Blythe and Castillo-Chavez 1989, Castillo-Chavez and Blythe 1989, and Castillo-Chavez *et al.* 1990). A constant ϕ-matrix implies that the preference structure in the population remains unchanged, but it is important to note that the values of p will change with time as a result of the dynamics of the model changing the sizes of the mixing sub-populations, $\{T_i(t)\}$, *i.e.* time-dependent changes in the set of mixing probabilities $p_{ij}(t)$ are not necessarily an indication of change in behavior. Even where all of the n groups in the population are identical (*i.e.* all the $C_i(t)$ are equal, and all the $T_i(t)$ are equal, so that $\rho_j(t) = 1/n$ for all j), the preference structure ϕ can produce non-random mixing in p (*c.f.* Blythe *et al.*, 1991*b*).

To illustrate the flexibility of the approach, we apply the framework in the context of a "classical" epidemic model for the homosexual transmission of gonorrhea among n interacting sub-populations, using as few definitions and equations as possible.

Let $S_i(t)$ and $I_i(t)$ respectively denote the number of susceptibles and infecteds in the i^{th} group, at time t. Let Λ_i denote the rate of influx (recruitment) of new susceptibles to the i^{th} group, and let $1/\mu$ and $1/\sigma$ be the average duration of a sexual "lifetime" and the average duration of the infected phase, respectively. If $B_i(t)$ is the incidence rate (of infections) in the i^{th} group, then we may write

$$\frac{dS_i(t)}{dt} = \Lambda_i - B_i(t) - \mu S_i(t) + \sigma I_i(t) \tag{20}$$

$$\frac{dI_i(t)}{dt} = B_i(t) - (\mu+\sigma)\, I_i(t) \tag{21}$$

for i = 1,..., n. We of course require initial conditions $S_i(0) > 0$, $I_i(0) \geq 0$ for all i. The incidence rates are given by

$$B_i(t) = S_i(t) \sum_{j=1}^{n} \lambda\, p_{ij}(t) C_i(t) \frac{I_j(t)}{T_j(t)}, \tag{22}$$

which is interpreted as follows. Each individual in group i has $C_i(t)$ partners per unit time, at time t. Of these, a fraction $p_{ij}(t)$ come from group j (j = 1,..., n), and of these in turn, a fraction $I_j(t)/T_j(t)$ are infected at time t ($T_j(t) = S_j(t) + I_j(t)$, ie., the total population of the i^{th} group). In this simple illustrative example, there is assumed to be a constant probability λ of a susceptible person becoming infected during a partnership with an infected person. Thus the summation term on the RHS of Eq (22) is the probability per unit time of a susceptible person in group i becoming infected at time t, and hence $B_i(t)$ is the total rate of new infections occurring in group i at time t. In this simplified form, if we prescribe the p matrix, then we have a complete specification of the gonorrhea epidemic model. See Blythe *et al.* (1991*b*) for more information on the formulation and implementation of mixing models in STD epidemics, and Blythe *et al* (1991*a*) for a discussion of possible methods for estimating the ϕ-values from the restricted data available.

(b) Another area where sexual mixing models can be of some importance is population genetics. As a first approximation at applying the generalized mixing framework in this area, consider the case of a recessive gene at a single locus, with no population regulation or frequency dependence (Blythe *et al.* 1991*b*). A good example (Crow 1986, pp. 50-53) is red-headedness in a human population. We divide the population into three groups: Group 1, homozygous (*AA*) individuals who do not carry the "red" allele; Group 2, heterozygotes (*Aa*) who carry but do not express the allele; and Group 3, homozygotes (*aa*) who express red hair. We assume that the fractions of offspring born from the six possible crosses follows the standard elementary random pattern (see *op. cit.*).

Clearly no one can distinguish *AA* from *Aa* phenotypes, so the only reasonable form for the preference or affinity matrix is

$$\phi = \begin{bmatrix} \alpha & \alpha & \beta \\ \alpha & \alpha & \beta \\ \beta & \beta & \gamma \end{bmatrix} \tag{23}$$

where $0 \leq \alpha, \beta, \gamma \leq 1$ are constants. Individuals in group 1 and 2 have the same preferences for *aa* versus non-*aa* (β versus α), and group 3 individuals have preferencies β and γ for non-*aa* and *aa* individuals, respectively. Assortative mating is represented by $\beta = \alpha = 0$, and $\gamma = r/\rho_3(t)$ where $r < 1$, *i.e.* a fixed fraction of the population of *aa* individuals mate among themselves, regardless of group population sizes.

In Blythe *et al.* (1991*b*) it is shown how a simple model may be derived under the following standard simplifying assumptions: (*a*) every individual in generation t has just one partner (mate); (*b*) the unit of time is the generation; (*c*) individuals from generation t are not counted in generation t + 1; (*d*) all matings produce 2ζ offspring ($\zeta>0$); and (*e*) the *a* allele frequency q remains constant. Then, $C_i(t) = 1$, for all i and t, and

$$\rho_i(t) = \frac{T_i(t)}{T(t)} \equiv x_i(t) \quad \text{for all} \quad i, \text{ and } T(t) = \sum_{k=1}^{n} T_k(t), \tag{24}$$

where T(t) is the total population, and $x_i(t)$ is the proportion of group i in the population, in generation t. Note that $x_1(t) + x_2(t) + x_3(t) = 1$ and $\frac{1}{2}x_2(t) + x_3(t) = q$.

Now for convenience write $Z_t \equiv x_3(t)$ as the proportion of *aa* individuals in the population in generation t. Then we have the recurrence relation

$$Z_{t+1} = f(Z_t) + g(Z_t)^2 / h(Z_t), \qquad Z_0 < q \tag{25}$$

where

$$f(Z) = \alpha q^2 - 2(\alpha-\beta)qZ + (\alpha-2\beta+\gamma)Z^2 ,$$

$$g(Z) = (1-\alpha)q + (\alpha-\beta)(1+q)Z - (\alpha-2\beta+\gamma)Z^2 ,$$

and

$$h(Z) = (1-\alpha) + (\alpha-\beta)Z - (\alpha-2\beta+\gamma)Z^2.$$

Blythe *et al.* (1991*b*) consider some of the properties of this map, and suggest directions in which future work on the impact of the new formalism in population genetics amy be determined.

Example 4: Joint Distributions.

Eq (5) seems also to be a new result in the theory of distributions. Consider X and Y, two jointly discrete random variables, with density function $f_{X,Y}(\cdot,\cdot)$, and marginal density functions $f_X(\cdot)$ and $f_Y(\cdot)$, respectively. If the joint density function is jointly symmetric, *i.e.* $f_{X,Y}(u,v) = f_{X,Y}(v,u)\ \forall\ (x,y)$ in the appropriate space, then clearly $f_X(u) = f_Y(u)$, $\forall$ u, so X and Y have the same function for their marginals. But these marginals are just ρ_u, and $\pi_{uv} \equiv f_{X,Y}(u,v)$, so that Eq (5) gives us a representation of all jointly symmetric discrete joint density functions.

In fact there is nothing new in the use of infinite families of joint distributions: for example Mood *et al.* (1974, p.142) use one to illustrate the fact that knowing the marginals does not necessarily mean that we know the joint distribution. Using discrete random variables, and a symmetric density function, we may write Mood *et al.*'s (1974) example as

$$\pi_{ij} = \rho_i \rho_j \left[1 + \nu \left(\Upsilon_i - Q\right) \left(\Upsilon_j - Q\right) \right], \tag{26}$$

where $-1 \leq \nu \leq +1$ is an arbitrary parameter,

$$\Upsilon_i \equiv \sum_{k=1}^{i} \rho_k , \tag{27}$$

is the cumulative distribution of the marginals, and

$$Q \equiv \sum_{m=1}^{n} \rho_k \Upsilon_k \tag{28}$$

(with continuous variables, $Q = \frac{1}{2}$). This example was originally due to D. Morgenstern (Plackett, 1965)), and has been extended and studied by *e.g.* Farlie (1960) and Gumbel (1960). Clearly we always have the same marginals, ρ, regardless of the value of ν, and of course the constraints Eqs (1) – (4) are satisfied. This means, incidently, that the p_{ij} deriving from Eq (26) could be used as a mixing function in STD modelling, as in *Example 2* above. What the new representation result says is that *all* such examples may be expressed in the form of Eq (5), it being required only to find the appropriate ϕ. It is easy to show that in Mood *et al*'s (1974) example, the choice

$$\phi_{ij} = \frac{\nu}{1+\nu Q^2} \Upsilon_i \Upsilon_j \tag{29}$$

in Eq (5) leads directly to the family Eq (26) as a special sub-class, for $0 \leq \nu \leq +1$. For the negative – and perfectly permissible – range of ν, Eq (29) leads to negative ϕ_{ij}, which is not permissible under the representation thereom. A set of values of ϕ which are non-negative may be obtained, however, by using

$$\phi_{ij} = 1 + \nu \left(\Upsilon_i - Q\right) \left(\Upsilon_j - Q\right) , \tag{30}$$

which works, with non-negative ϕ, for the full range of ν. In this case, all the $R_i = 0$, and we require that the limit given in Eq (A7) be defined. The relaxation of the assumption that the ϕ are strictly

constants is important only if the marginals change with time, a point discussed in Blythe *et al.* (1991*a*,*b*).

This is a rather good example for making the point that the relationship between ϕ and π is not unique (which was, after all, essentially the point that Mood *et al.* (1974) were making). However, because Eq (5) is a representation of *all* π, this non-uniqueness need no longer be regarded as something of a *bête noire*, as we show in the following example.

The extension to continuous variables is trivial, regaining the original formulation of Busenberg and Castillo-Chavez (1989, 1990). Consider X and Y, two jointly continuous random variables, with jointly symmetric density function $\pi(x,y)$, and hence marginal density functions $\rho(x)$ and $\rho(y)$. Then π is related to ρ by an equation exactly analogous to Eq (5), with $R(\cdot)$ now defined in an integral rather than a summation sense, and the denominator likewise. For example, if $\rho(x) = \frac{1}{h}$ for $x \in [0,h]$ and zero elsewhere, *i.e.* uniformly distributed marginals, and say

$$\phi(x,y) = \frac{1}{2h}(x+y), \quad (x,y) \in [0,h], \tag{31}$$

then we have directly that

$$\pi(x,y) = \frac{1}{16h^2}\left[2+7\frac{xy}{h^2}+\left[4-\frac{x}{h}\right]\left[4-\frac{y}{h}\right]\right], \quad (x,y) \in [0,h]\,. \tag{32}$$

This is a convenient way of generating joint density functions with a specified "affinity" structure $\phi(x,y)$, but the reverse process of finding $\phi(x,y)$ such that a given $\pi(x,y)$ is obtained is often difficult. For example, say $\pi(x,y)$ is the bivariate normal distribution

$$\pi(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\xi^2}} \cdot \exp\left[-\frac{1}{2\sigma^2(1-\xi^2)}\left[(x-\mu)^2 - \xi(x-\mu)(y-\mu) + (y-\mu)^2\right]\right], \tag{33}$$

for $(x,y) \in (-\infty, +\infty)$, with μ and σ^2 the mean and variance, respectively, of the marginal $\rho(\cdot)$, and ξ the correlation coefficient of X and Y. Clearly,

$$\rho(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{(z-\mu)^2}{2\sigma^2} \right] . \tag{34}$$

To find suitable $\phi(x,y)$ we have to equate

$$\Phi(x,y) \equiv \exp\left[-\tfrac{1}{2}\ln(1-\xi^2) - \frac{\xi}{2\sigma^2(1-\xi^2)} \left[\xi(x-y)^2 - 2(1-\xi)(x-\mu)(y-\mu) \right] \right] \tag{35}$$

with $\Psi(x,y)$ (the continuous version of the function in Appendix A), and solve for $\phi(x,y)$. An obvious solution is $\phi(x,y) = \Psi(x,y) = \Phi(x,y)$, which is not very informative, but which may be the only one.

If we wish to generate an alternative symmetric joint distribution with marginals f(u) suct that $U \sim N(\mu, \sigma^2)$, we may easily do so by choosing $\phi(x,y)$ with properties reflecting our assumptions about affinity. Say we choose to parameterize ϕ as

$$\phi(x,y) = \begin{cases} a \geq 0, & \text{if } x = y, \\ b \geq 0, & \text{if } x \neq y \end{cases} , \tag{36}$$

which is a pronounced version of the "like-with-like" functions studied by *e.g.* Blythe and Castillo-Chavez (1989) and Castillo-Chavez *et al.* (1989), and know as "diagonal mixing" in the context of sexual mixing models (Blythe and Castillo-Chavez, 1991). Then we have at once that

$$f_{X,Y}(x,y) = f(x)f(y)\Big\{ K\Big(1-b-(a-b)f(x)\Big)\Big(1-b-(a-b)f(y)\Big) + a\delta(x-y) + b\Big(1-\delta(x-y)\Big)\Big\} , \tag{37}$$

where

$$K = \frac{2\sqrt{\pi}\,\sigma}{2\sqrt{\pi}\,\sigma(1-b)-a} , \tag{38}$$

and $\delta(\cdot)$ is the Dirac delta function. This joint density function is of course quite different from Eq 33), although the marginals are the same. For Eq (38) we find that the correlation coefficient is given

simply by

$$\xi = \frac{(a-b)}{2\sqrt{\pi}\,\sigma}, \tag{39}$$

which may be either positive or negative, depending on whether or not the diagonal parameter, a, dominates. When a = b Eq (37) reduces to $f_{X,Y}(x,y) = f(x)\,f(y)$, *i.e.* independent variables, and there is naturally zero correlation between X and Y.

In the above examples we have one or two parameters, varying over specified subranges, describing an infinite family of distributions, all of which have the same marginals. In these cases, these parameters (ν in Eq (30), ξ in Eq (35), a and b in Eq(37)) act to parameterize the affinity function ϕ (see also Farlie, 1960; Gumbel, 1960; Plackett, 1965; Johnson and Kotz, 1972). We may thus offer the interpretation that the ϕ provide a means for describing and implementing the most general form of correlation structure in symmetric bivariate distributions, and that their non-uniqueness with respect to the marginals is no more a restriction than it is with the correlation coefficient ξ in the bivariate normal case, subject only to the issues of parameter estimation. We are at present extending the results on symmetric joint distributions to the more general bivariate case where π need not be symmetric (this appears to be trivial: Blythe, MS in prep), and examining the implications for higher-dimension problems.

DISCUSSION AND CONCLUSIONS

We have presented a new result in the description of the probability of event-pairs where within-pair order is immaterial, Eq (5), indicated how the proof underlying the result is arrived at, and suggested four areas where useful applications may be found.

The underlying concept is that under the specified conditions for event-pair processes, we may make use of the concept of "affinity" between pairs of events of different types. The elements of the ϕ matrix characterize the degree to which particular types of events appear together in pairs in a non-random fashion. The card-game example (*Example 1*) provides a clear illustration of how the result,

and the ϕ-method, may be applied, but perhaps represents the least interesting area of application. The potential for the method to provide a consistent theoretical basis for the measurement of species association (*Example 2*) is perhaps one of the most exciting areas of potential, particularly if other, similar applications are found. Clearly, work is required both in the general analysis of the association-measure technique, and in the design of experimental protocols to make use of it. Applications in the areas of sexually transmitted disease dynamics and in population genetics (*Example 3*) are by now appreciated, and work is under way to develop these themes. The final application, *Example 4*, is particularly interesting, indicating as it does that any symmetric bivariate joint distribution may be represented as a member of the ϕ-family. In particular, this gives us a generalization of the idea of correlation structure, and provides a framework for the construction of bivariate distributions with specified properties given know marginals.

ACKNOWLEDGEMENTS

This research has been partially supported by NSF grant DMS-8906580, NIAID Grant R01 A129178-01, and Hatch project grant NYC 151-409, USDA to Carlos Castillo-Chavez. S. P. Blythe's research has been partially supported by funds from the office the dean of the College of Agriculture and Life Sciences at Cornell University, and the Wellcome Trust.

APPENDIX A

Busenberg and Castillo-Chavez (1989, 1990) developed their representation thereom in the context of sexual mixing patterns, where p_{ij} is known as a "mixing matrix". They used continuous variables for age and risk (*c.f.* Example 3, above), but in the context of this paper we need only consider discrete distributions, indexed in S. We consider only the result where no event-pairs are forbidden; the result for arbitrary permissible sub-sets does not differ in any major respect, although some restrctions on the ϕ may be necessary (*c.f.* Blythe 1991).

(*Busenberg and Castillo-Chavez 1990, Thereom 4.3*) Given the axiomatic framework of Eq (1) – (3) and (8),

$$p_{ij} = \rho_j \, \Psi_{ij} \tag{A1}$$

is a general solution: all p satisfying the axioms may be writen in this form, with Ψ constrained to be strictly positive in the sub-set $(i,j) \in S'$, jointly symmetric ($\Psi_{ji} = \Psi_{ij}$), and satisfying the contraint

$$\sum_{k \in S'} \rho_k \, \Psi_{ik} \; = \; 1, \quad i \in S' \tag{A2}$$

For $(i,j) \notin S'$ but $\in S$, we clearly have $p_{ij} = 0$, and Ψ_{ij} need not be defined. The proof that (A1) is a mixing function is trivial, amounting to showing that the axioms are satisfied. The proof of the converse, that all mixing functions can be written in this form, requires assuming p an arbitrary mixing function, and showing that defining $\Psi_{ij} = p_{ij}/\rho_j$ satisfies the axioms.

(*After Busenberg and Castillo-Chavez 1990, Thereom 4.4*) To proceed, we need to find a helpful representation of Ψ. Noting that Ψ is jointly symmetric and strictly positive in S', we can always find a set of real strictly positive numbers $\{\theta_i \; ; \; i \in S'\}$ such that $\Psi_{ij} \geq \theta_i \, \theta_j$, and hence we may always represent Ψ in the form

$$\Psi_{ij} = \theta_i \, \theta_j \; + \; \omega_{ij} \, , \tag{A3}$$

where $\omega_{ji} = \omega_{ij} \geq 0$, $(i,j) \in S'$. It is then a consequence of the constraint (A2) that we can write

$$\theta_i = \frac{\left[1 - \sum_{k \in S'} \rho_k \omega_{ik}\right]\left[1 - \sum_{k \in S'} \rho_k \omega_{ik}\right]}{\sqrt{\sum_l \rho_l \left[1 - \sum_{k \in S'} \rho_k \omega_{lk}\right]}} \quad , \quad i \in S' \ , \tag{A4}$$

i.e. the representation Eq (A3) of Ψ, given constraint (A2), means that θ may always be represented in the form (A4), as functions of ω. Substitution of (A4) into (A3) and then into (A1), and identifying ω with ϕ, reveals that p_{ij} is always representable in the form

$$p_{ij} = \rho_j \left[\frac{R_i R_j}{\sum_{m \in S'} \rho_m R_m} + \phi_{ij} \right] , (i,j) \in S'. \tag{A5}$$

For all θ_i to be real and non-negative, we require

$$\sum_{l \in S'} \rho_l \left[1 - \sum_{k \in S'} \rho_k \omega_{lk}\right] > 0 \ , \tag{A6}$$

although the product-pairs $\theta_i \theta_j$ will be real and non-negative even when (A5) is violated. Equality with zero in Eq (A6) is permissible only when *all* the R_i's equal zero, with the limit

$$\lim_{\substack{R_k \to 0 \\ k \in S'}} \left[\frac{R_i R_j}{\sum_{m \in S'} \rho_m R_m} \right] = 0 \tag{A7}$$

defined. Then we have a completely degenerate solution, in the sense defined by Blythe (1991), and $\Psi = \phi$.

Remark: We can also represent Ψ with respect to an upper bound rather than a lower bound, as in (A3). In that case we obtain a set of negative "disaffinities" $\sigma_{ij} = -\phi_{ij}$. This adds no new information, however.

REFERENCES

Anderson, R.M. and May, R.M. (1984) Spatial, temporal and genetic heterogeneity in host populations and the design of immunization programmes. *IMA J. Math. Appl. Med. & Biol.* 1, 233-266.

Anderson, R.M., Medley, G.F., May, R.M. and Johnson, A.M. (1987) A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J. Math. Appl. Med. &. Biol.* 3, 229-263

Anderson, R.M., Blythe, S.P., Gupta, S. and Konings, E. (1989) The transmission dynamics of the Human Immunodeficiency Virus Type 1 in the male homosexual community in the United Kingdom: the influence of changes in sexual behavior. *Phil. Trans. R. Soc. Lond.* B 325, 145-198

Anderson, R.M., Gupta, S. and Ng, W. (1990) The significance of sexual partner contact networks for the transmission dynamics of HIV. *J. A. I. D. S.* 3, 417-429

Barbour, A.D. (1978) MacDonald's model and the transmission of bilharzia. *Trans. Royal Soc. Trop. Med. Hyg.* 72, 6-15.

Blythe, S.P. (1991) Heterogeneous sexual mixing in populations with arbitrarily connected multiple groups (To appear in *Math. Pop. Stud.*).

Blythe, S.P. and Anderson, R.M. (1989) Heterogeneous sexual activity models of HIV transmission in male homosexual populations. *IMA J. Math. Appl. Med. & Biol.* 5, 237-260

Blythe, S.P. and Castillo-Chavez, C. (1989) Like-with-like preference and sexual mixing models. *Math. Biosci.* 96, 221-238.

Blythe, S.P. and Castillo-Chavez, C. (1990*a*) The one-sex mixing problem: a choice of solutions? (Submitted *J. of Math. Biol.*)

Blythe, S.P. and Castillo-Chavez, C. (1990*b*) Like-with-like mixing and sexually transmitted disease epidemics in one-sex populations (Manuscript.)

Blythe, S.P., Castillo-Chavez, C. and Casella, G. (1991*a*) Empirical methods for the estimation of the mixing probabilities for socially structured populations from a single survey sample. (Submitted *Math. Pop. Studies.*)

Blythe, S.P., Castillo-Chavez, C., Palmer, J. and Cheng, M. (1991*b*) Towards a unified theory of mixing and pair formation. (Submitted to *Math. Biosci.*)

Busenberg, S. and Castillo-Chavez, C. (1989) Interaction, pair formation and force of infection terms in sexually transmitted diseases. *Lect. Notes Biomath.* 83, 289-300

Busenberg, S. and Castillo-Chavez, C. (1990) A general solution of the problem of mixing of subpopulations, and, its application to risk- and age- structured epidemic models. (To appear in *IMA J. Math. Appl. Med. & Biol.*)

Castillo-Chavez, C. and Blythe, S.P. (1989) Mixing framework for social/sexual behavior. *Lect. Notes Biomath.* 83, 275-288

Castillo-Chavez, C. and Busenberg, S. (1990) On the solution of the two-sex problem. In: *Proceedings of the Interantional Conference on Differential Equations and Applications to Biology and Population Dynamics* (S. Busenberg and M. Martelli, eds.), Lecture Notes in Biomathematics, Springer-Verlag (in press).

Crow, J.F. (1986) ***Basic concepts in quantitative and evolutionary genetics.*** W.H. Freeman & Co., New York

Dietz, K. and Schenzle, D. (1985) Proportionate mixing models for age-dependent infection transmission. *J. Math. Biol.* 22, 117-120.

Farlie, D.J.G. (1960) The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* 47, 307-323

Gumbel, E.J. (1960) Bivariate exponential distributions. *J. Amer. Stat. Assoc.* 55,698-707

Gupta, S., Anderson, R.M. and May, R.M. (1989) Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS* **3**, 1-11

Hethcote, H.W. and Yorke, J.A. (1984) *Gonorrhea transmission dynamics and control. Lect. Notes Biomath.* **56**

Hubalek, Z. (1973) Coefficients of association and similarity based on arbitrary (presence-absence) data. *Biol. Rev.*, **57**, 669-689

Hyman, J.M. and Stanley, E.A. (1989) The effect of social mixing patterns on the spread of AIDS. *Lect. Notes Biomath.* 81, 190-219

Jacquez, J.A., Simon, C.P., Koopman, J., Sattenspiel, L. and Perry, T. (1988) Modelling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.* **92**, 119-199.

Jacquez, J.A., Simon, C.P. and Koopman, J. (1989) Structured mixing: heterogeneous mixing by the definition of mixing groups. *Lect. Notes Biomath.* **83**, 301-315.

Johnson, N.L. and Kotz, S. (1972) "DISTRIBUTIONS IN STATISTICS, VOL IV: Continuous Multivariate Distributions" *John Wiley & Sons, Inc.* New York

Koopman, J.S., Simon, C.P., Jacquez, J.A., Joseph, J., Sattenspiel, L. and Tae Sung Park (1988) Sexual partner selectiveness on homosexual HIV transmission dynamics. *J. A. I. D. S.* **1**, 486-504

Koopman, J.S., Simon, C.P., Jacquez, J.A. and Tae Sung Park (1989) Selective contact within structured mixing with an application to HIV transmission risk from oral and anal sex. *Lect. Notes Biomath.* 83, 316-348.

Krebs, C.J. (1978, 2nd Edition) "ECOLOGY: The Experimental analysis of Distribution and Abundance". *Harper & Row*, New York

Legendre, L. and Legendre, P. (1983) "NUMERICAL ECOLOGY". *Elsevier.* New York

Ludwig, J.A. and Reynolds, J.F. (1988) "STATISTICAL ECOLOGY". *Wiley Interscience,* New York

Mielke, P.W. and Siddiqui, M.M. (1965) A combinatorial test for independence of dichotomous responses. *J. Amer. Stat. Assoc.* **60**, 437-441

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974, 3rd Edition) "INTRODUCTION TO THE THEORY OF STATISTICS". *McGraw-Hill*, New York

Nold, A. (1980) Heterogeneity in disease-transmission modeling. *Math. Biosci.* **52**, 227–240.

Plackett, R.L. (1965) A class of bivariate distributions. *J. Amer. Stat. Assoc.* **60**, 516-522

Wolff, R.W. (1989) "STOCHASTIC MODELING AND THE THEORY OF QUEUES". *Prentice-Hall*, New Jersey.

27/12/90